Statistical Machine Learning Methods for Bioinformatics
II. Hidden Markov Model for Biological Sequences

Jianlin Cheng, PhD

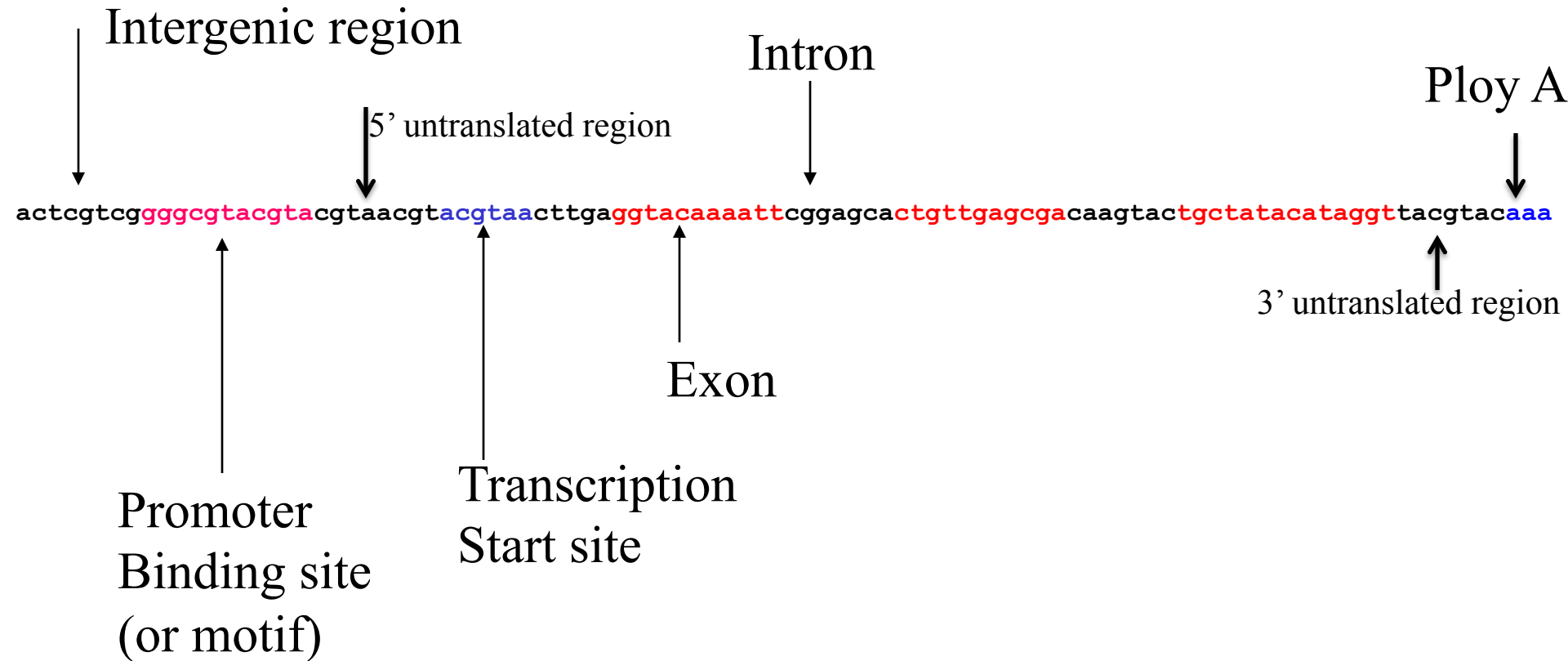Department of Computer Science & Informatics Institute

University of Missouri

2012

Application of HMM in Biological Sequence Analysis

- Gene prediction
- **Protein sequence modeling (learning, profile)**
- **Protein sequence alignment (decoding)**
- **Protein database search (scoring, e.g. fold recognition)**
- Protein structure prediction
- …

Motif and Gene Structure

Intergenic region

5' untranslated region

Intron

Ploy A

actcgtcggggcgtacgtacgtaacgtacgtaacttgaggtacaaaattcggagcactgttgagcgacaagtactgctatacataggttacgtacaaa

3' untranslated region

Promoter
Binding site
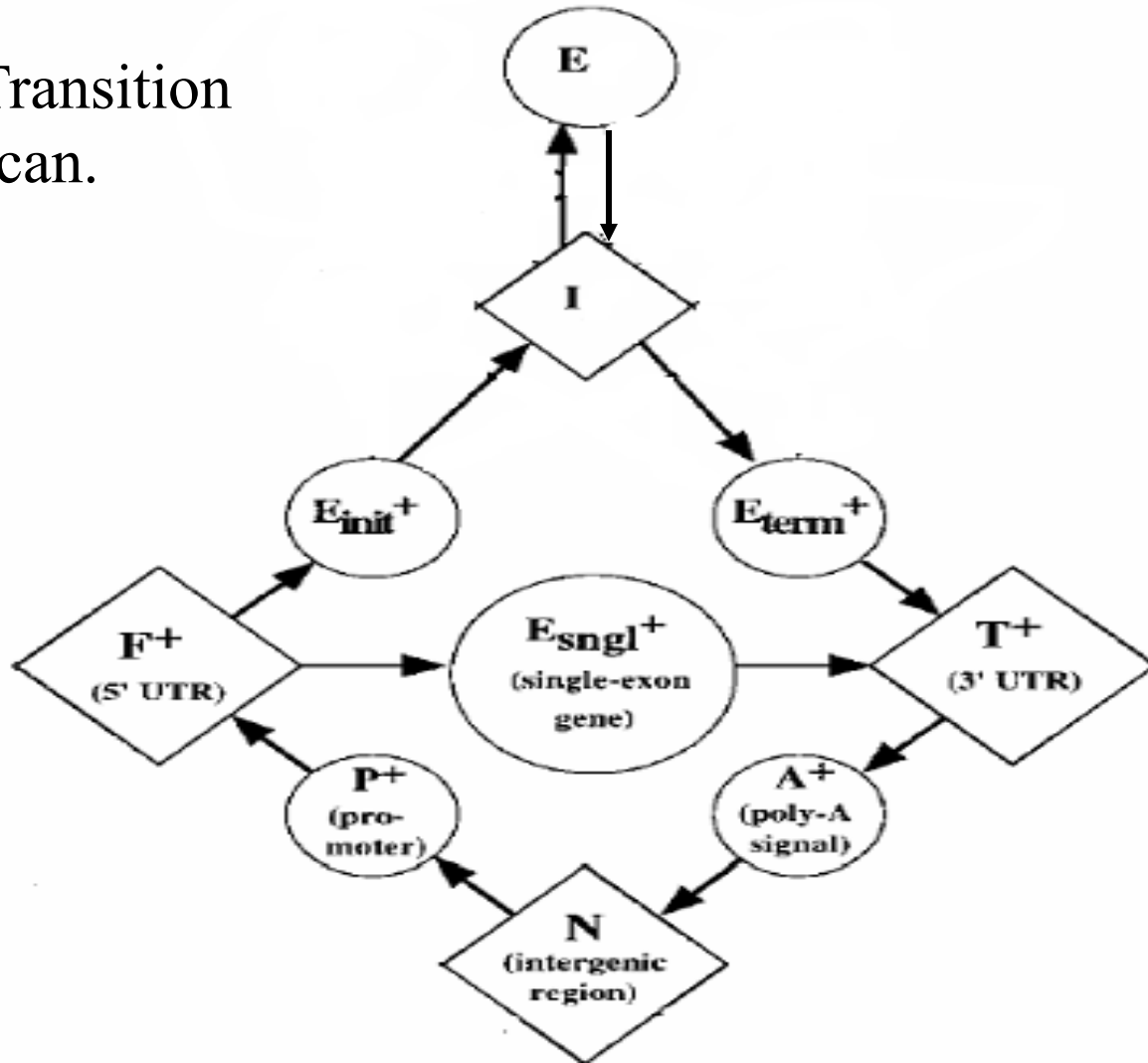(or motif)

Transcription
Start site

Exon

HMM has been used for modeling binding site
and gene structure prediction.

GENSCAN

(genes.mit.edu/GENSCAN.html)

Simplified State Transition
Diagram of GenScan.

Model Protein Family (Profile HMM)

- Create a statistical model (HMM) for a group of related protein sequences (e.g protein family)

- Identify core (conserved) elements of homologous sequences

- Positional evolutionary information (e.g. insertion and deletion)

Example: Hemoglobin Transports Oxygen

Heterocyclic Ring

Iron atom

Conserved Position
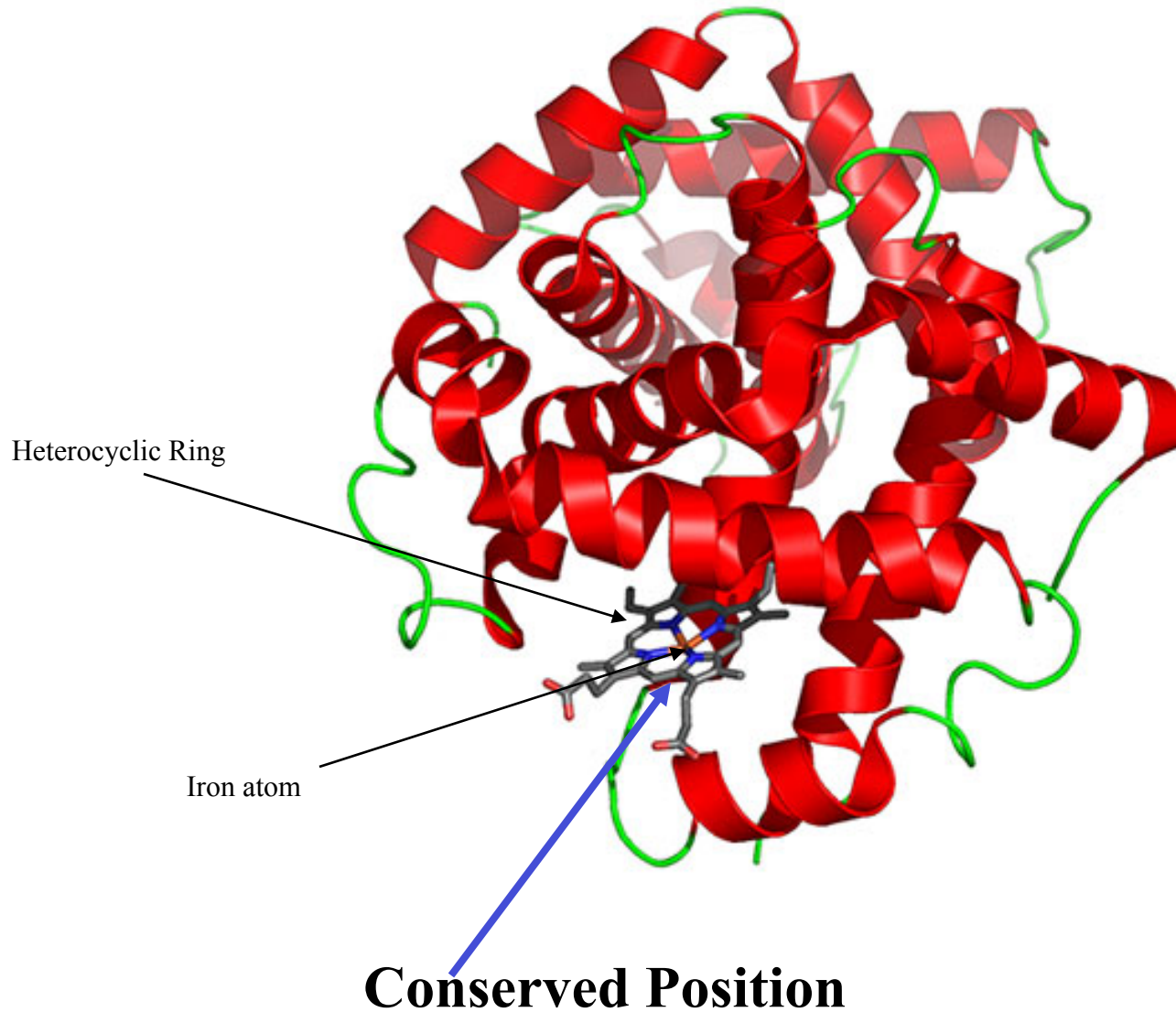
Why do We Build a Profile (Model)?

- Understand the conservation (core function and structure elements) and variation
- Sequence generation
- Multiple sequence alignments
- Profile-sequence alignment (more sensitive than sequence-sequence alignment)
- Family / fold recognition
- Profile-profile alignment

Protein Family

```
seq1   VRRNNMGMPLIESSSSYHDALFTLGYAGDRISQMLGMRLLAQGRLSEMAGADALDV
seq2   NIYIDSNGIAHIYANNLHDLFLAEGYYEASQRLFEIELFGLAMGNLSSWVGAKALSS
seq3   SAETYRDAWGIPHLRADTPHELARAQGTARDRAWQLEVERHRAQGTSASFLGPEALSW
seq4   DRLGVVTIDAANQLDAMRALGYAQERYFEMDLMRRAPAGELSELFGAKAVDL
```

```
seq1   ---VRRNNMGMPLIESSSSYHDALFTLGY--AGDRISQMLGMRLLAQGRLSEMAGADALDV
seq2   --NIYIDSNGIAHIYANNLHDLFLAEGYYEASQRLFEIELFG-LAMGNLSSWVGAKALSS
seq3   SAETYRDAWGIPHLRADTPHELARAQGT--ARDRAWQLEVERHRAQGTSASFLGPEALSW
seq4   ------DRLGVVTIDAANQLDAMRALGY--AQERYFEMDLMRRAPAGELSELFGAKAVDL
```

Imagine these sequences evolve from a single ancestral sequence
and undergo evolutionary mutations. How to use a HMM to model?

Key to Build a HMM is to Set Up States

- Think about the positions of the ancestral sequence is undergoing mutation events to generate new sequences in difference species. A position can be modeled by a **dice**.

- **Match** (match or mutate): the position is kept with or without variations / mutations.

- **Delete**: the position is deleted

- **Insert**: amino acids are inserted between two positions.
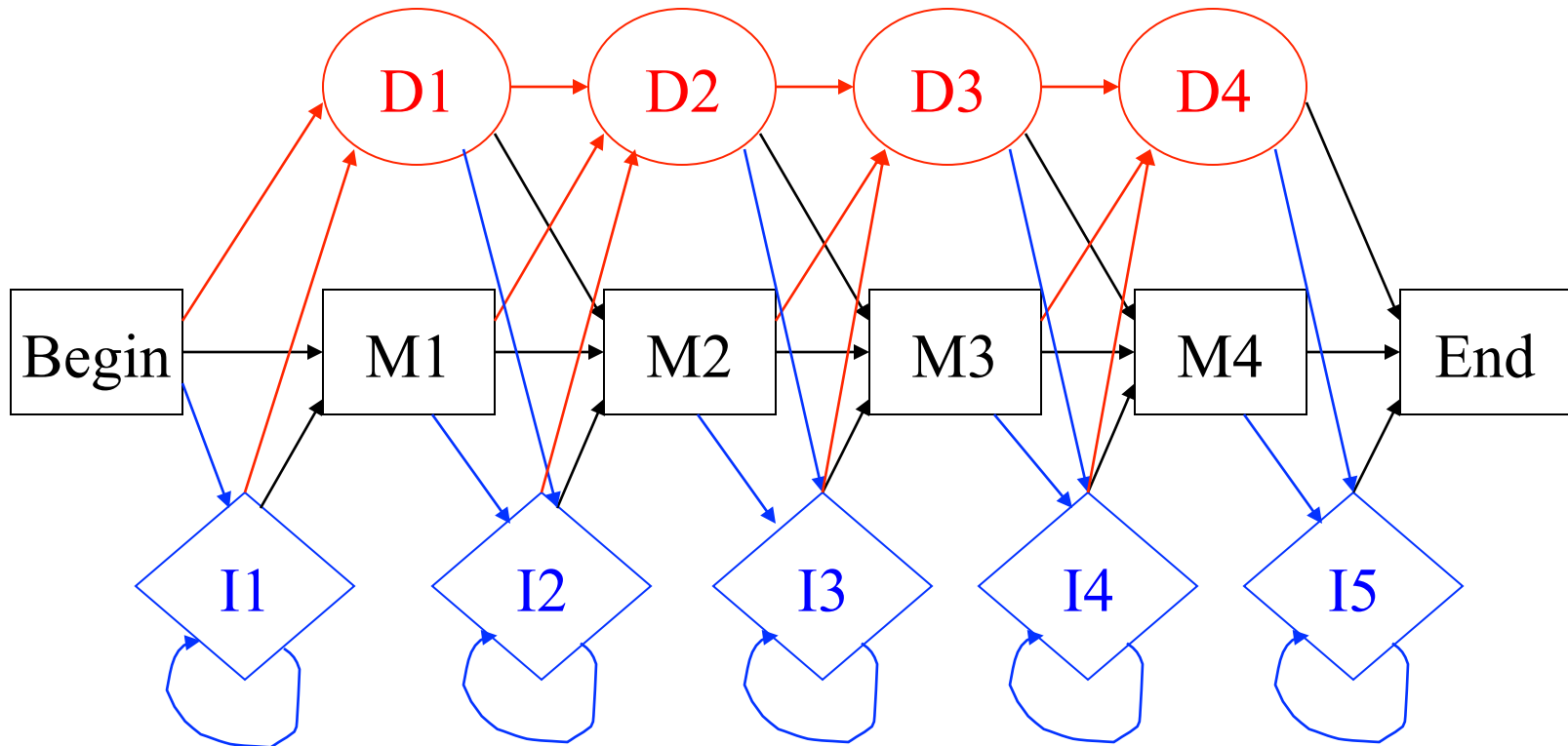
Hidden Markov Model


```
Begin → M1 → M2 → M3 → M4 → End
```
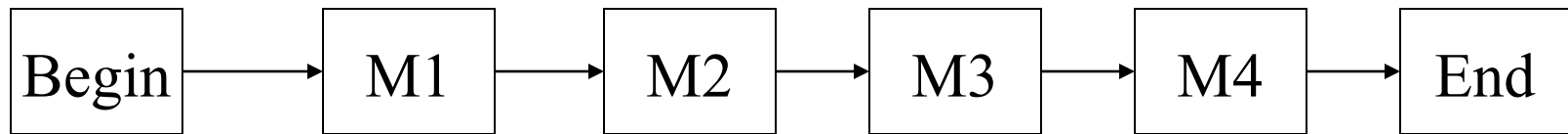
Each match state has an emission distribution of 20 amino acids; one match state for a position.

Krogh et al. 1994, Baldi et al. 1994

Hidden Markov Model
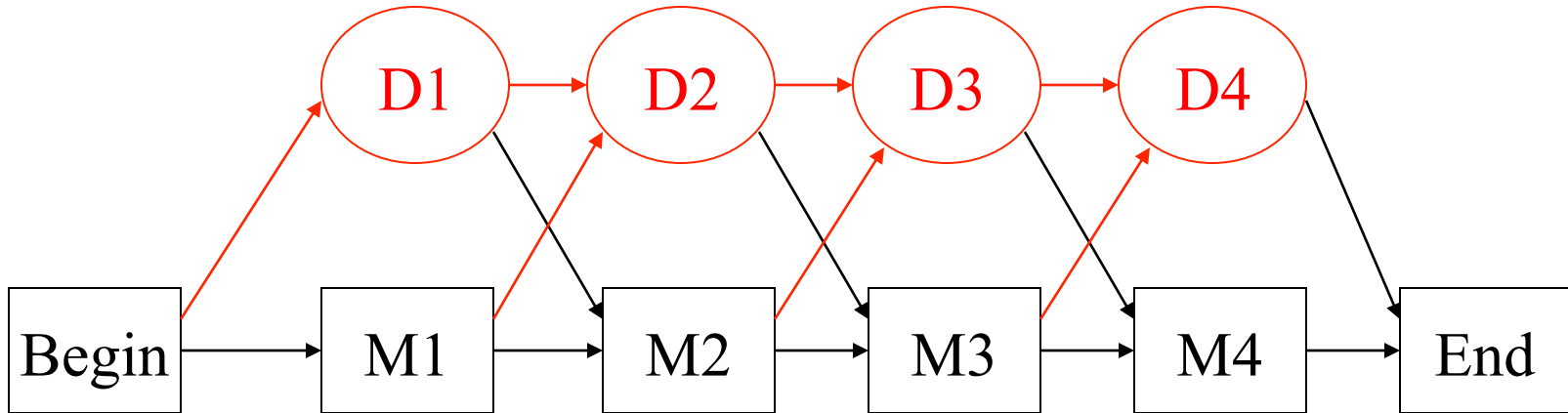
Each match state has an emission distribution of 20 amino acids.
Deletion state is a mute state (emitting a dummy)

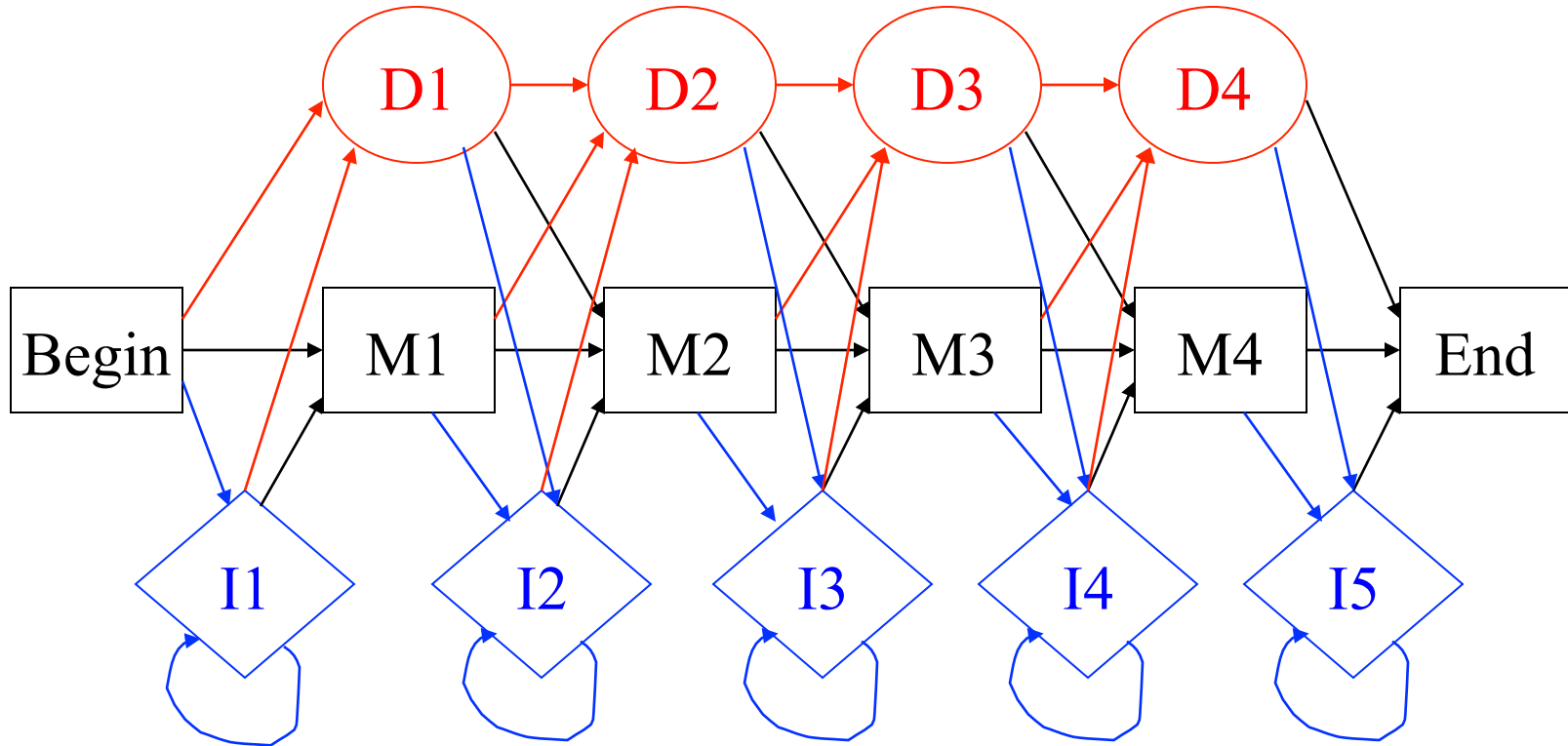Krogh et al. 1994, Baldi et al. 1994

Hidden Markov Model

Krogh et al. 1994, Baldi et al. 1994

Each match state has an emission distribution of 20 amino acids.
Each insertion state has an emission distribution of 20 amino acids.
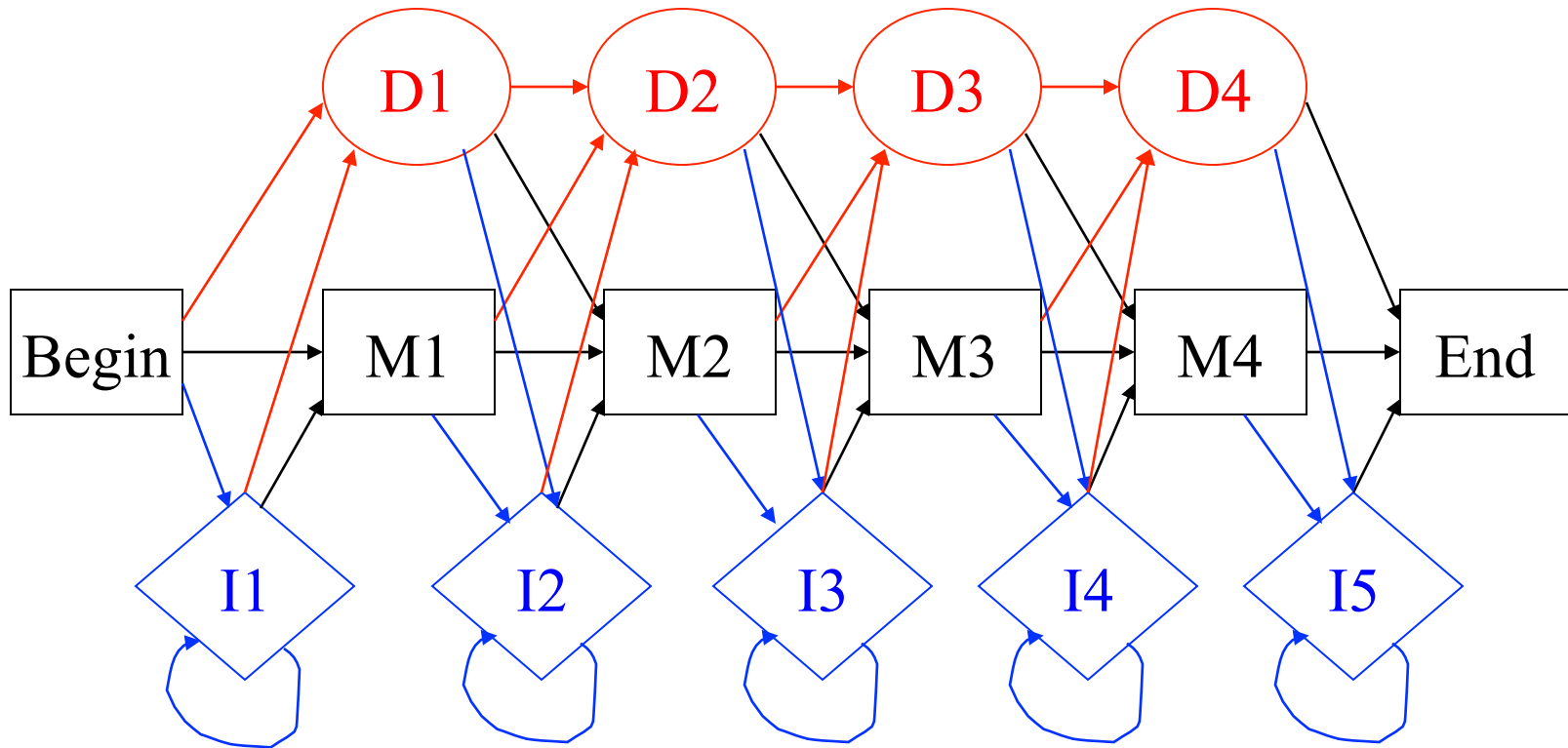Variants of architecture exist. (see Eddy, bioinformatics, 1997)

Hidden Markov Model

How many states? (M positions: length of model)

M (match) + M (deletion) + (M+1) (insertion) + 2 = 3M + 3
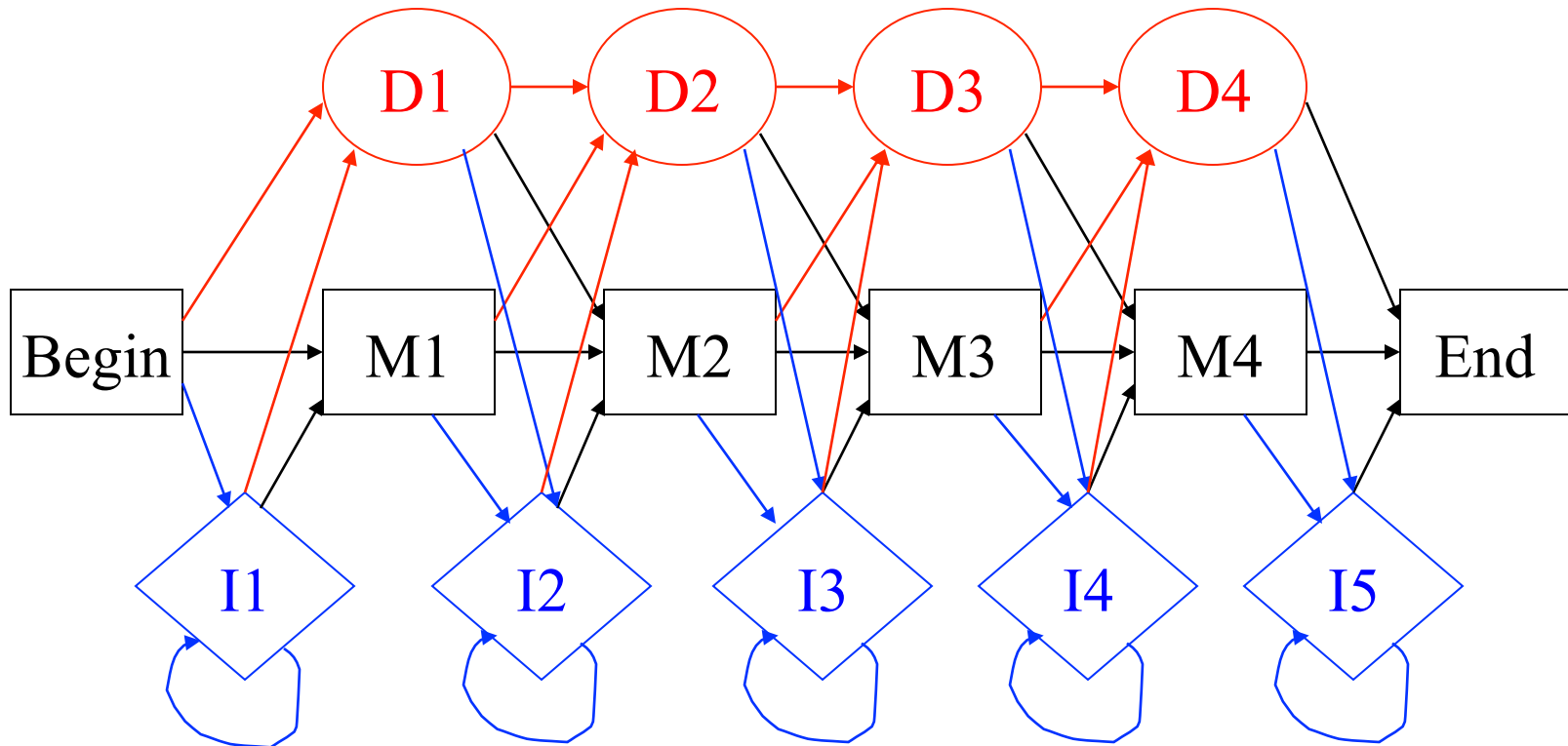
Krogh et al. 1994, Baldi et al. 1994

Hidden Markov Model

How many transitions? (M positions: length of model)
Deletion: $3M - 1$, Match: $3M - 1$, Insertion: $3(M+1) - 1$, B/E: 3
Total $= 9M + 3$.

Krogh et al. 1994, Baldi et al. 1994

Hidden Markov Model

How many emissions? (M positions: length of model)
M * 20 (match) + (M+1)*20 (insertion) = 40M + 20

Krogh et al. 1994, Baldi et al. 1994

Initialization of HMM

- How to decide model length (the number of match states)?

- How to initialize transition probabilities?

- How to initialize emission probabilities?

How to Decide Model Length?

- **Learn**: Use a range of model length (centered at the average sequence length). If transition probability from a match (M_i) state to a delete state (D_{i+1}) > 0.5, remove the M_{i+1}. If transition probability from a match (M_i) state to an insertion state (I_{i+1}) > 0.5, add a match state.

- **Get from multiple alignment**: assign a match state to any column with <50% gaps.

How to Initialize Parameters?

- Uniform initialization of transition probabilities is ok in most cases.

- Uniform initialization of emission probability of insert state is ok in many cases.

- Uniform initialization of emission probability of match state is **bad**. (lead to bad local minima)

- Using amino acid distribution to initialize the emission probabilities is better. (need regularization / smoothing to avoid zero)

Initialize from Multiple Alignments

```
seq1  ---VRRNNMGMPLIESSSSYHDALFTLGY--AGDRISQMLGMRLLAQGRLSEMAGADALDV
seq2  --NIYIDSNGIAHIYANNLHDLFLAEGYYEASQRLFEIELFG-LAMGNLSSWVGAKALSS
seq3  SAETYRDAWGIPHLRADTPHELARAQGT--ARDRAWQLEVERHRAQGTSASFLGPEALSW
seq4  -------DRLGVVTIDAANQLDAMRALGY--AQERYFEMDLMRRAPAGELSELFGAKAVDL
```

First, assign match / main states, delete states, insert states from MSA
Get the path of each sequence
Count the amino acid frequencies emitted from match or insert states, which are converted into probabilities for each state (need smoothing/ regularization / pseudo-count).
Count the number of state transitions and use them to initialize transition probabilities.

Estimate Parameters (Learning)

- We want to find a set of parameters to maximize the probability of the observed sequences in the family: maximum likelihood: P(sequences | model) = P(sequence 1| model) * … * P(sequence n | model).

- Baum-Welch's algorithm (or EM algorithm) (see my previous lectures about HMM theory)

Demo of HMMER (learning)

Visualization of Features and Structure in HMM
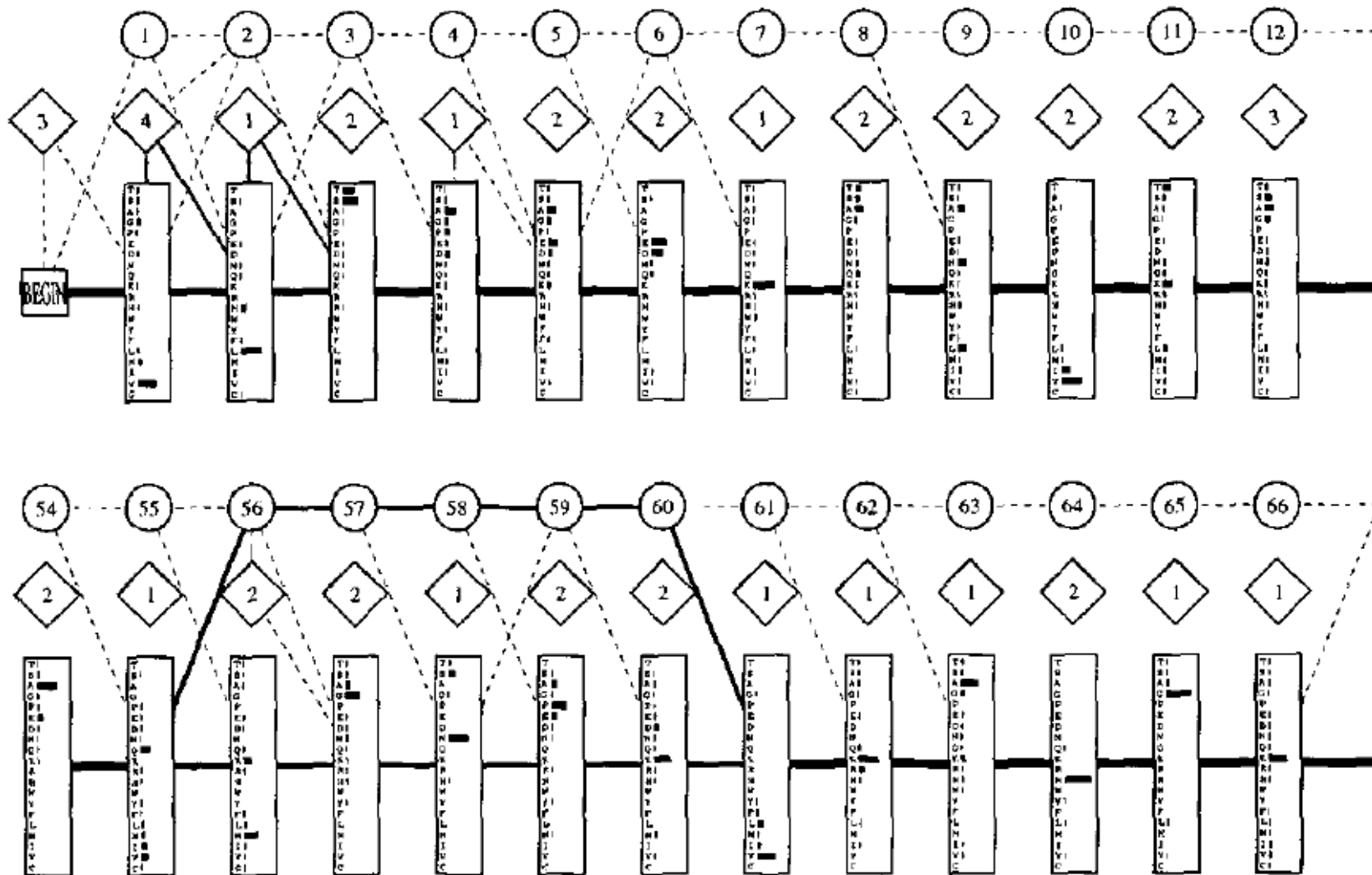
Myoglobin protein family. How to interpret it?

Krogh et al. 94

Demo of HMMEditor
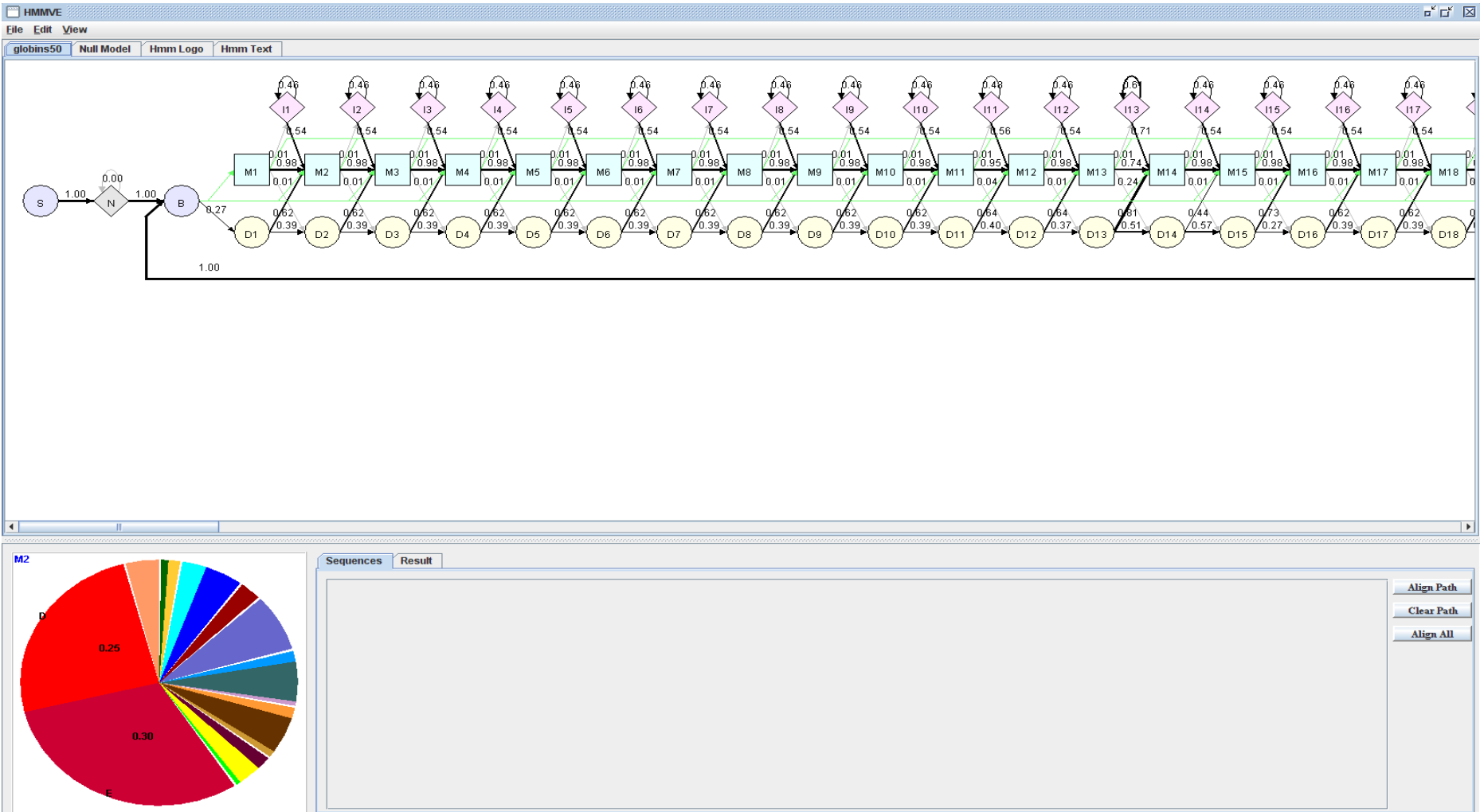
J. Dai and J. Cheng. **HMMEditor: a visual editing tool for profile hidden Markov model**. *BMC Genomics*, in press, 2007.
Tool: http://casp.rnet.missouri.edu/hmmeditor/ **(data in /mlbioinfo/, 1MSB)**
Paper: http://www.biomedcentral.com/qc/1471-2164/9/S1/S8

Protein Family Profile HMM Databases

- Pfam database (Sonnhammer et al., 1997, 1998) (domain database) (http://pfam.wustl.edu/)

- PROSITE profiles database (Bairoch et al., 1997) (motif database)

What Can We Do With the HMM?

- **Recognition and classification. Widely used for database search**: does a new sequence belong to the family? (database search)

- **Idea**: The sequences belonging to the family (or generated from HMM) should receive higher probability than the sequence not belong to the family (unrelated sequences).

Two Ways to Search

- Build a HMM for each family in the database. Search a query sequence against the database of HMMs. (Pfam)

- Build a HMM for a query family, and search HMM against of a database of sequences

Compute P(Sequence | HMM)

- Forward algorithm to compute P(sequence | model)

- We work on: -log(P(sequence|M): distance from the sequence to the model. (negative log likelihood score)

- Unfortunately, -log(P(sequence|M) is length dependent. So what can we do?

Normalize the Score into Z-score

- Search the profile against a large database such as Swiss-Prot
- Plot –log(P(seuqence|model), NLL scores, against sequence length.
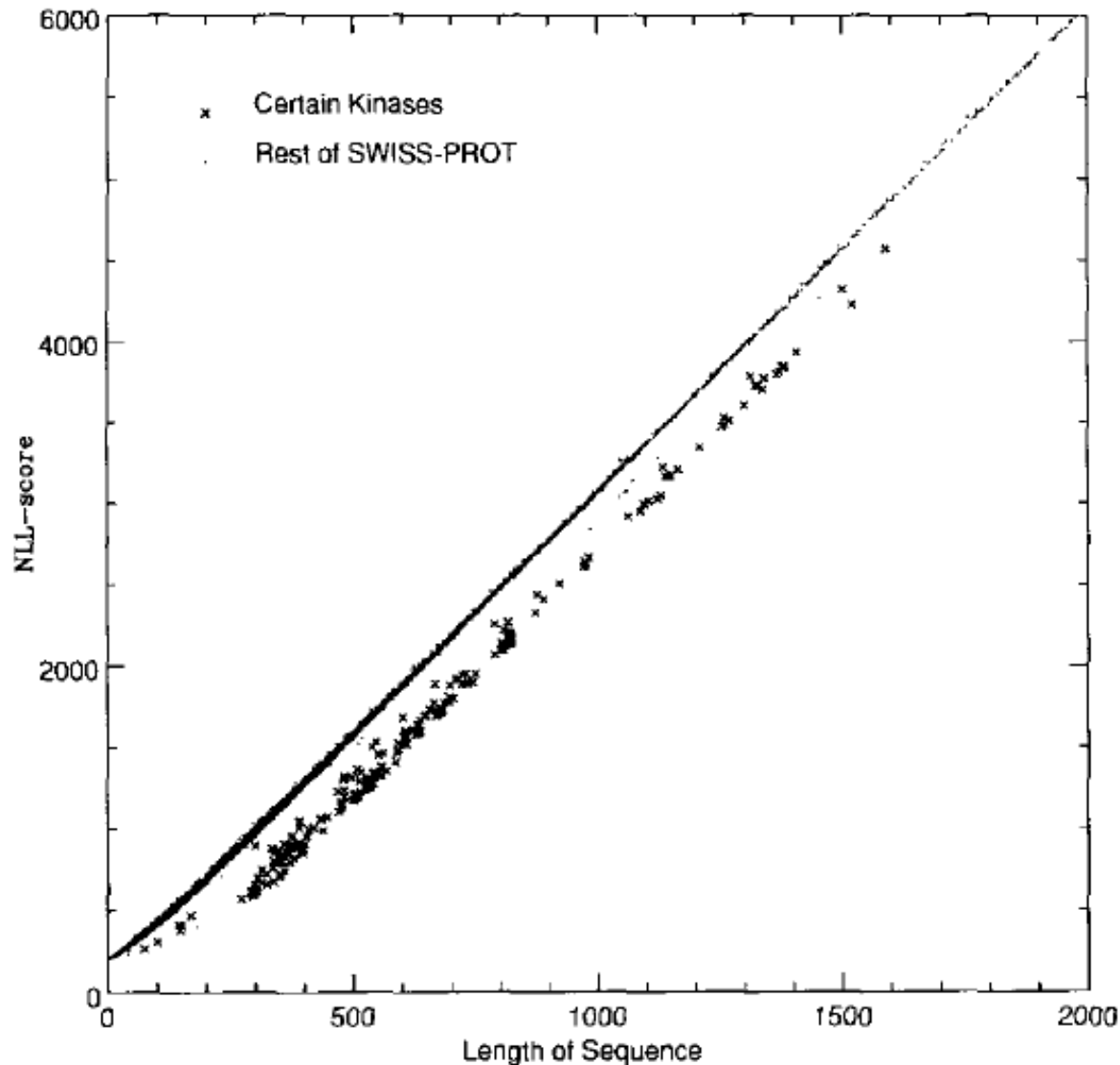
Figure 9. Scatter plot of NLL-score versus length for sequences in SWISS-PROT using the Kinase HMM.

NLL score is linear to sequence length.
NLL scores of the same family is lower than un-related sequences
We need normalization.
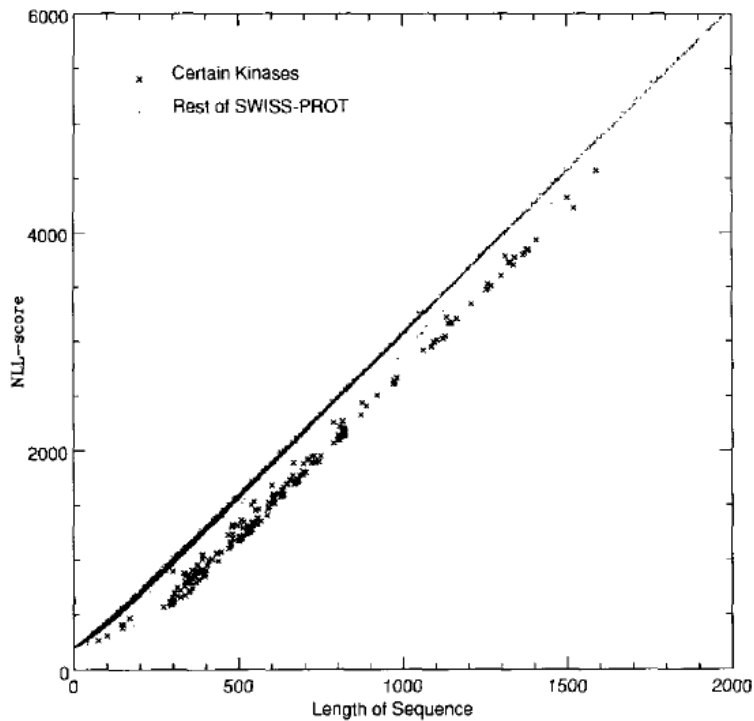
Krogh et al. 94

NULL model of unrelated sequences:

Figure 9. Scatter plot of NLL-score versus length for sequences in SWISS-PROT using the Kinase HMM.

Length	Mean (u)	Std (σ)
100	500	5
101	550	6

Compute Z-score: $|s - u| / \sigma$

Z > 4: the sequence is very different from unrelated sequence.
(for non-database search, a randomization can work.)

Extreme Value Distribution
(Karlin and Altschul)

Log-odds score = log(P(seq|λ) / P(seq |null))

$$P(S \geq x) = 1 - \exp(-Kmn\ e^{-\lambda x})$$

P-value

E-value
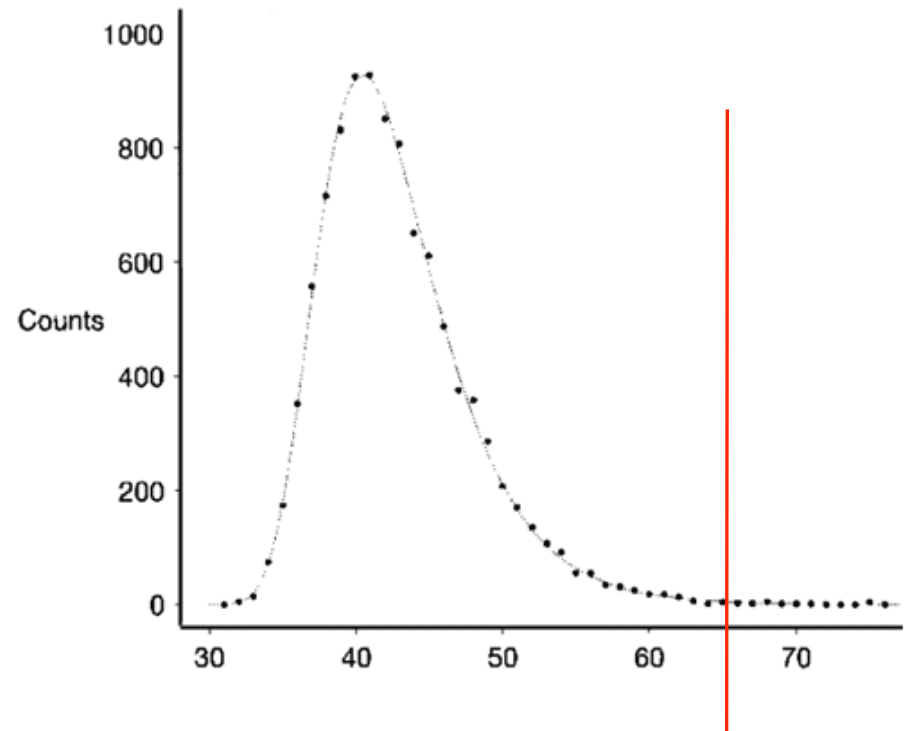
K and lamda are statistical parameters. m,n model and sequence length.

HMMer search demo (casp.rnet.missouri.edu/mlbioinfo)

Project I: Soybean Protein Classification

- 2000 Transcription Factors (proteins) in Arabidopsis

- These sequences are clustered into 64 families by biologists (known)

- 4000 new proteins in Soybean?

- How to assign them into those families?

- Do you need to use a multiple sequence alignment tool?

How to solve the problem?

http://casp.rnet.missouri.edu/mlbioinfo/soybean_data/

Iterative Data Search (SAM)

- Use BLAST to search a query sequence database to gather an initial MSA
- **Repeat**

  Build an HMM from MSA (training)

  Search the HMM against sequence database to get more related sequences

  Create a new MSA from all related sequences
- **Until** a predefined number of iterations or no new sequences are found.

(Idea is very similar as PSI-BLAST)

Insight I: Evaluating a Sequence Against a Profile HMM is Fold Recognition Process or Function Prediction

- Check if a sequence is in the same family (or superfamily) as the protein family (superfamily) used to build the profile HMM.

- If they are in the same family, they will share the similar protein structure (fold), possibly protein function.

- The known structure can be used to model the structure of the proteins without known structure.

- The known function can be used to predict function.

Insight II: Evaluating a Sequence Against a HMM is Sequence-Profile Alignment

- Align a query sequence against a HMM of the target sequence to get the most likely path (Viterbi algorithm) (or vice versa)

- Match the path of the query sequence with the path of the target sequence, we get their alignment.

- Represented work: SAM or HMMER.

Pairwise Alignment via HMM

Seq 1: A T G R   K E
Path : M_1 I_1 I_1 M_2 D_3 M_4 I_4

Seq 2: V C   K   E   R   P
Path : M_1 I_1 M_2 M_3 M_4 I_4

Path: M_1 M_2 M_3 M_4
Seq 1: A T G R   -   K   E
Seq 2: V   C - K   E   R   P

HMM for Multiple Sequence Alignment

- Build a HMM for a group of sequences
- Align each sequence against HMM using Viterbi algorithm to find the most likely path. (dynamic programming)
- Match the main/match states of these paths together.
- Add gaps for delete states
- For insertion between two positions, use the longest insertion of a sequence as template. Add gaps to other sequence if necessary. (see Krogh's paper)

Demo for Multiple Sequence Alignment Using HMMEditor

How About Evaluating the Similarity between HMMs?

- Can we evaluate the similarity of two HMMs?

- Can we align two profile HMMs? (profile-profile alignment). Compare HMM with HMM.

J. Soeding, Bioinformatics, 2005

HMM-HMM Comparison Profile-Profile Alignment
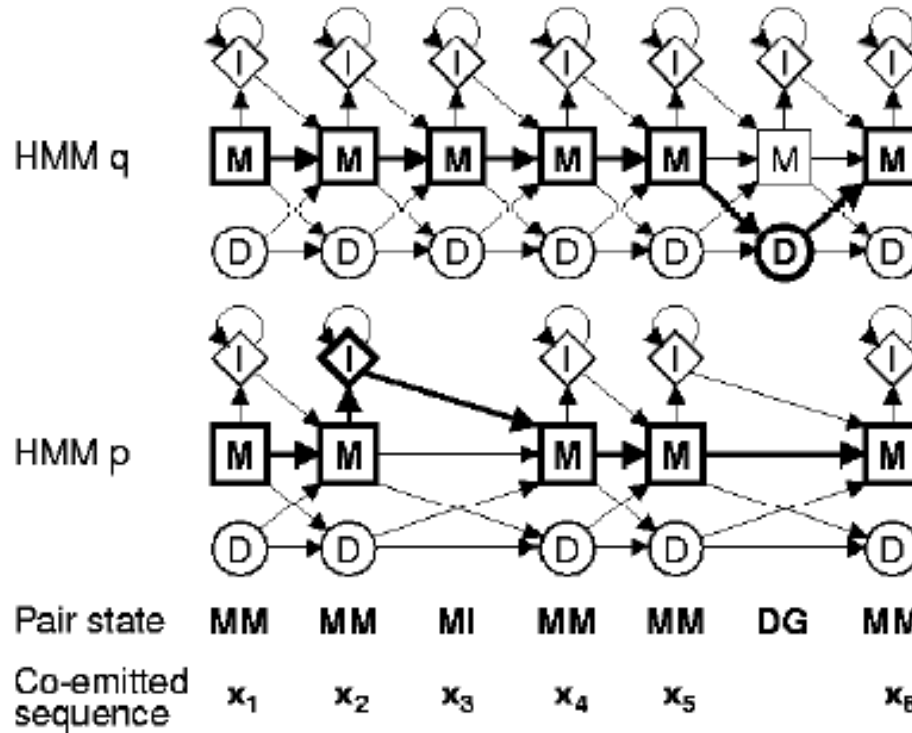
Generalized Sequence (of States) 1

Generalized Sequence (of States) 2
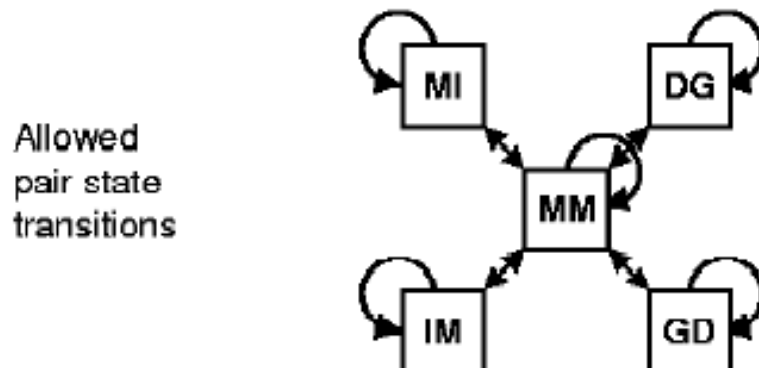
Goal: find a sequence of paired states with Maximum sum of log odds scores

J. Soeding, Bioinformatics, 2005

COACH Approach

- Given two families of sequences, build a multiple alignment (MSA) for each one of them.

- Build HMM from one MSA

- Align another MSA against the HMM. (match each column of amino acids against states in the HMM)

R. C. Edgar and K. Sjolander. Bioinformatics, 2004

How to Do Local Alignment

- COACH approach

  With respect to sequence: add an insertion state right after the start state and right before the end state.

  With respect to HMM: start state can jump to any match state and any match state can jump to end state.

Sequence Weighting

- **Henikoff-Henikoff**: sum of position-based weight. For each position, each type of amino acid is assigned weight 1. The weight is of each amino acid is 1 / frequency of the amino acid at the position. The weight of a sequence is the sum of positional weights. (gap is not counted. A position with more than 50% gaps may be removed from counting.) An easy, useful algorithm
- **Tree algorithm**: construct a phylogenetic tree. Start from root, weight 1 flows down. At any branch, the weight is cut to half.

Pseudo-Count

- ## PSI-BLAST pseudo-count
(Altschul et al., 1997)

For each position of PSSM, score is log (Q_i / P_i). Q_i is the estimated probability of residue i to be found in the column. P_i is the background probability.

Small sample size require prior knowledge of residue i to better estimate Q_i. Best method is Dirichlet Mixtures. A very good and simple method is a data-dependent pseudocount method. This method uses the prior knowledge of amino acid relationships embodied in the substitution matrix S_{ij} to generate residue pseudocount frequencies g_i. (f_j is the frequency of residue j). α is set to Nc -1 (Nc is the number of columns). β is set to 10.

$$g_i = \sum_j \frac{f_j}{P_j} q_{ij}$$

$$s_{ij} = [\ln(q_{ij}/P_i P_j)]/\lambda_u$$

$$q_{ij} = P_i P_j e^{\lambda_u s_{ij}}$$

$$Q_i = \frac{\alpha f_i + \beta g_i}{\alpha + \beta}$$

Null Model

- Background null model (log-odds)
- Reverse null model (SAM)
- Sum of log-odds score of local alignment obeys extreme-value distribution (same as PSI-BLAST): good for estimating the significance of sequence-HMM match.
- Sum of log-odds score of global alignment is length dependent.

HMM Software and Code

- HMMER: http://hmmer.wustl.edu
- SAM: http://www.cse.ucsc.edu/research/combio/sam.html
- HHSearch: http://toolkit.tuebingen.mpg.de
- PRC-HMM: http://supfam.mrc-lmb.cam.ac.uk/PRC/
- COACH: http://www.drive5.com/lobster/
- HMM-HMM comparison: http://www.brics.dk/~cstorm/hmmcomp/
- MUSCLE: http://www.drive5.com/muscle/

Project 2: Multiple Sequence Alignment Using HMM

- Dataset: BaliBASE: *http://www-bio3d-igbmc.u-strasbg.fr/balibase/*
- Generate multiple alignment using Clustalw (http://www.ch.embnet.org/software/ClustalW.html) for a family of sequences
- Construct a HMM for a family of sequences (initialization, number of states)
- Estimate the parameters of HMM using the sequences
- Analyze the emission probability of states (visualization)
- Analyze the transition probability between states (visualization)
- Compute the probability of each sequence
- Generate multiple sequence alignments and compare it with the initial multiple alignment
- Implement your own HMM or use open source code.

Reference: A. Krogh et al, JMB, 1994 and open source HMM.
J. D. Thompson, F. Plewniak and O. Poch. Bioinformatics, 1999